Comparaison de méthodes de classification de données fonctionnelles

Boubacar Diallo, Ndèye Niang, Vincent Audigier, Feriel Bouhadjera

Comparaison de méthodes de classification de données fonctionnelles

Boubacar Diallo[*,**], Ndèye Niang[*,**]
Vincent Audigier[*,**] Feriel Bouhadjera[*,**]

[*]Laboratoire CEDRIC, équipe MSDMA, 2 rue Conté, 75003 Paris
boubacar.diallo8@lecnam.net
https://cedric.cnam.fr/equipes/msdma
[**]Conservatoire National des Arts et Métiers, 2 rue Conté, 75003 Paris

Résumé. En classification de données fonctionnelles par des approches basées sur les distances, la partition obtenue est sensible au choix de ces dernières. Plusieurs mesures de distance ont été proposées dans la littérature. Nous étudions ici l'impact sur la partition de l'intégration de la dérivée première dans le calcul des distances en termes d'indice de qualité interne (Silhouette) et externe (Indice de Rand ajusté). Pour cela, nous menons tout d'abord une étude comparative de plusieurs de ces approches à travers différents scénarios de simulation de données fonctionnelles où les groupes diffèrent par la forme des courbes. Par la suite, cette étude est complétée par une évaluation sur la base de différents jeux de données réelles. Les résultats montrent que l'intégration de la dérivée peut avoir un effet important sur la qualité de la partition en fonction de la nature des différences entre groupes.

1 Introduction

La classification basée sur des approches géométriques repose sur la définition d'une distance, ou parfois seulement d'une dissimilarité, permettant de formaliser la notion de ressemblance entre objets. Dans le cas particulier de la classification sur données fonctionnelles, cette dernière porte sur des courbes (fonctions) qui relève d'un choix de distance spécifique. La distance D_0, extension de la distance euclidienne aux fonctions, est la plus classiquement employée.

Elle mesure la ressemblance entre les courbes en terme d'amplitude verticale sur le domaine de définition des fonctions. Toutefois, quand les courbes diffèrent par d'autres variations que leur amplitude, cette distance n'est alors plus adaptée. Dans ce contexte, il peut alors être pertinent d'utiliser la distance D_1 qui revient à calculer la distance D_0 sur les fonctions dérivées. Ainsi, on retrouve dans Meng et al. (2018) une proposition d'algorithme de k-means fonctionnel s'appuyant sur la combinaison de ces deux distances. Plus récemment, Yu et al. (2025) propose d'utiliser une approche par filtrage où deux Analyse en Composantes Principales Fonctionnelles (ACPF) sont effectuées : une première sur les fonctions et une seconde sur les fonctions dérivées. La distance

entre courbes est définie à partir des composantes principales de ces deux analyses, permettant de prendre en compte à la fois les variations en amplitude sur les fonctions et sur leurs dérivées avec un système de pondération. Ces distances établies, de nombreux algorithmes de classification géométrique peuvent alors être appliqués directement. On peut par exemple citer le k-means, ou la Classification Ascendante Hiérarchique (CAH). Toutefois, on retrouve également dans la littérature des algorithmes plus spécifiques aux données fonctionnelles (e.g. Zhou et al., 2023).

L'objet de ce travail est d'étudier l'impact de la prise en compte de la dérivée première sur la classification à travers une étude comparative. Dans la suite, nous présentons d'abord le prétraitement des données fonctionnelles, plus précisément le lissage, qui constitue une étape incontournable de l'analyse des données fonctionnelles (Section 2). Puis, nous aborderons les méthodes de classification de données fonctionnelles, en revenant notamment sur la définition des distances (Section 3). Enfin, nous comparerons ces différents méthodes sur la base d'une étude de simulation (Section 4) et de données réelles fonctionnelles classiquement utilisées (Section 5).

2 Données fonctionnelles

En raison des limitations techniques des instruments de mesure en matière d'enregistrement et de capacité de stockage, les données fonctionnelles sont généralement observées de manière discrétisée. Afin de se ramener à des données continues, le traitement des données fonctionnelles nécessite une opération de lissage (e.g. Ramsay et Silverman, 2005). Par ailleurs, l'analyse de ces données lissées passe souvent par une approche dite par filtrage, consistant à se ramener à des données tabulaires sur lesquelles il est facile d'appliquer des méthodes d'analyse classiques. L'ACPF est une technique commune pour s'y ramener. Nous revenons ci-dessous sur ces deux techniques usuelles.

2.1 Lissage

Soit n courbes, nous supposons que le i-ième individu est observé sur un ensemble de p mesures $\{y_{i1}, \ldots, y_{ip}\}$ prise à des points finis distincts $\{t_1, \ldots, t_p\}$ dans un intervalle $T \subset \mathbb{R}$, et contaminées par des bruits, selon le modèle :

$$y_{ij} = f_i(t_j) + \epsilon_{ij}, \quad i = 1, \ldots, n$$

où $f_i \in \mathcal{L}^2(T)$, est la vraie fonction (inconnue), de carré intégrable sur l'intervalle T, pour le i-ème individu et ϵ_{ij} sont des variables aléatoires indépendantes et identiquement distribuées de moyenne 0 et de variance finie σ^2.
En supposant que la fonction f_i soit une combinaison linéaire de fonctions de base $(\varphi_s)_{1 \leq s \leq d}$:

$$f_i(t) = \sum_{s=1}^{d} \theta_{is} \varphi_s(t), \quad t \in T,$$

les coefficients (θ_{is}) pour $1 \leq i \leq n$, $1 \leq s \leq d$ sont classiquement obtenus en résolvant le problème d'optimisation suivant :

$$\underset{\boldsymbol{\theta}_i \in \mathbb{R}^d}{\arg\min} \, \|\mathbf{y}_i - \Phi \boldsymbol{\theta}_i\|^2 + \lambda_i \, \boldsymbol{\theta}_i^\top \mathbf{P} \boldsymbol{\theta}_i, \tag{1}$$

où, pour un i fixé, $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})^\top$ est le vecteur d'observations, $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{id})^\top$ vecteur de coefficients, $\lambda_i \geq 0$ désigne le paramètre de pénalisation du lissage. La matrice Φ de dimension $p \times d$ dont les colonnes représentent les fonctions de base $(\varphi_s(t_1), \ldots, \varphi_s(t_p))$ pour $s = 1, \ldots, d$, et $\mathbf{P}$ est la matrice de pénalisation définie par $p_{sl} = \int_T \varphi_s^{(2)}(t)\varphi_l^{(2)}(t)dt$, pour $s, l = 1, \ldots, d$, où $\varphi_s^{(2)}$ désigne la dérivée seconde de la fonction de base φ_s. On note que si $\lambda_i = 0$ dans (1), on parlera de lissage par Moindres Carrés Ordinaires (MCO) ; sinon ($\lambda_i > 0$), on parlera de lissage par Moindres Carrés Pénalisés (MCP). La solution de ce problème de minimisation est donnée par :

$$\hat{\boldsymbol{\theta}}_i = (\Phi^\top \Phi + \lambda_i \mathbf{P})^{-1} \Phi^\top \mathbf{y}_i. \tag{2}$$

Ainsi, la fonction lissée du i-ème individu est donnée par :

$$\hat{f}_i(t) = \hat{\boldsymbol{\theta}}_i \Phi^\top,$$

où les coefficients $\hat{\boldsymbol{\theta}}_i$ sont donnés par (2).
Le lissage MCP offre une meilleure capacité à retrouver la fonction sous-jacente. De plus, il donne une grande liberté dans le choix du nombre de fonctions de base, indépendamment du nombre de mesures, ce qui peut s'avérer intéressant dans certains cas d'études. Cependant le choix du paramètre de pénalité λ_i est crucial et est souvent effectué à l'aide d'une Validation Croisée Généralisée (VCG), afin d'obtenir le meilleur équilibre entre la fidélité des données (risque de sur-ajustement) et la perte d'informations (sous-ajustement).

2.2 Analyse en composantes principales fonctionnelles

Cette méthode repose sur la décomposition de Kosambi–Karhunen–Loève où chaque fonction $\{f_i(t),\, t \in T\}$ est représentée par :

$$f_i(t) = \mu(t) + \sum_{s=1}^{K} \xi_{is}\, \phi_s(t), \tag{3}$$

où $\mu(t)$ désigne la fonction moyenne, $(\xi_{i1}, \ldots, \xi_{iK})$ représente le vecteur des scores et $(\phi_1(t), \ldots, \phi_K(t))$ la base de fonctions orthogonales. On définit ainsi l'opérateur de covariance empirique par :

$$(\Gamma_n f)(s) = \int_T C_n(s,t) f(t) dt,$$

où,

$$C_n(s,t) = \frac{1}{n} \sum_{i=1}^{n} \big(f_i(s) - \mu(s)\big)\big(f_i(t) - \mu(t)\big),$$

désigne la fonction de covariance empirique. L'opérateur de covariance est diagonalisé pour obtenir les fonctions propres, notées $\{(\hat{\phi}_s(t))_{1 \leq s \leq K},\, t \in T\}$, correspondant chacune à un mode principal de variation, ainsi que leurs valeurs propres, notées $(\hat{\lambda}_s)_{1 \leq s \leq K}$, correspondant à la variance expliquée par chacune des composantes. Les scores $(\hat{\xi}_{is})_{1 \leq s \leq K}$ obtenus pour chaque individu i sont les représentations discrètes de la i-ème courbe.

3 Méthodes de classification

Nous abordons ici les distances qui peuvent être employées pour intégrer l'information de la dérivée première dans une classification géométrique, puis présentons différentes méthodes de classification les prenant en compte.

3.1 Distances pondérées

Pour intégrer l'information issue de la dérivée première dans le processus de classification, deux approches peuvent être employées. D'une part, une transformation des données fonctionnelles en représentation vectorielle via une ACPF (voir section 2.2) peut être opérée en parallèle sur les fonctions lissées $\left(\hat{f}_i\right)_{1\leq i\leq n}$ et leurs dérivées $\left(\hat{f}_i^{(1)}\right)_{1\leq i\leq n}$. On obtient ainsi, les scores $\left(\hat{\xi}_{is}\right)_{1\leq s\leq K_0}$ pour chaque observation i, correspondant à $\hat{f}_i$ et les scores $\left(\hat{\xi}_{is}^{(1)}\right)_{1\leq s\leq K_1}$ pour chaque observation i, correspondant à $\hat{f}_i^{(1)}$. Ainsi, la distance pondérée entre deux fonctions s'écrit :

$$\tilde{D}_\omega(f_i,f_j)=\left\{(1-\omega)\sqrt{\sum_{s=1}^{K_0}\left(\hat{\xi}_{is}-\hat{\xi}_{js}\right)^2}+\omega\sqrt{\sum_{s=1}^{K_1}\left(\hat{\xi}_{is}^{(1)}-\hat{\xi}_{js}^{(1)}\right)^2}\right\}^{1/2}, \tag{4}$$

où $\omega\in[0,1]$.

D'autre part, une fois les courbes reconstruites via le lissage (voir section 2.1), on peut calculer la similarité, donnée par la distance pondérée fonctionnelle :

$$\begin{aligned} D_\omega(f_i,f_j)&=\sqrt{(1-\omega)\int_T\left(f_i(t)-f_j(t)\right)^2dt+\omega\int_T\left(f_i^{(1)}(t)-f_j^{(1)}(t)\right)^2dt}\\ &=\sqrt{(1-\omega)\,D_0^2(f_i,f_j)+\omega\,D_1^2(f_i,f_j)}. \end{aligned} \tag{5}$$

où, $i,j\in\{1,\ldots,n\}$, $f_i^{(1)}$ est la dérivée première de la fonction f_i. Les intégrales présentes dans cette expression peuvent être approchées numériquement par une approximation de Riemann.

Dans ces deux types de distance, ω est un hyperparamètre (optimisé par *gridsearch*) qui permet de régler l'importance des dérivées vis à vis des fonctions elles mêmes. Dans le cas particulier où $\omega=0$, la distance indiquée en équation 5 est équivalente à la distance D_0, tandis que si $\omega=1$, cette dernière revient à la distance D_1.

3.2 Méthodes

Les distances pondérées fournissent deux matrices de dissimilarité (4) et (5) entre les courbes sur lesquelles les méthodes classiques de classification peuvent être appliquées. Nous appliquons la CAH avec une stratégie de *Ward*, qui minimise l'augmentation de la variance intra-groupe. Au terme du processus d'agrégation (toutes les courbes sont dans

un unique groupe), un dendrogramme est obtenu. La meilleure partition en G classes est obtenue en cherchant un coude dans le diagramme en barres représentant les indices de niveaux d'agrégation associés au dendrogramme. Nous considérons également un k-means, dans laquelle la distance entre les individus et les centroïdes est définie à partir des distances pondérées (4) et (5). Enfin, sur les matrices de dissimilarité, il est possible d'appliquer l'algorithme PAM (Partitionning Around Medoids) comme dans Kaufman et Rousseeuw (2009) qui, à la différence du k-means, met à jour le représentant de chaque classe comme la courbe (médoïde) minimisant la distance totale aux individus de sa classe.

On trouve dans Yu et al. (2025), une autre approche appelée, *Subspace Projected Functional classification algorithm with weighted distance* ($SPFC_{\tilde{D}_\omega}$). Elle consiste en une opération de réaffectation (*Leave-One-Curve-Out*) des courbes partant d'une partition initiale obtenue à partir de la distance (4). Pour chaque courbe i exclue, la procédure met à jour les courbes moyennes $\hat{\mu}_{-i}^{(C)}(t)$ et leurs dérivées, puis recalcule les K_C composantes principales fonctionnelles $\hat{\phi}_{s,-i}^{(C)}(t)$ et les scores associés $\hat{\xi}_{is}^{(C)}$. Pour chaque groupe C, la courbe f_i est prédite par :

$$\hat{f}_{i(C)}^{K_C}(t) = \hat{\mu}_{-i}^{(C)}(t) + \sum_{s=1}^{K_C} \hat{\xi}_{is}^{(C)}\, \hat{\phi}_{s,-i}^{(C)}(t),$$

et de même pour sa dérivée. L'affectation optimale est obtenue en minimisant la distance pondérée $\tilde{D}_\omega$ entre la courbe observée et sa prédiction :

$$\hat{C}^*(f_i) = \arg\min_{C \in \mathcal{P}} \tilde{D}_\omega^{K_C}\left(f_i, \hat{f}_{i(C)}^{K_C}\right)^2.$$

On répète les étapes jusqu'à ce qu'aucune courbe ne change plus de groupe.

D'autres travaux, comme ceux de Zhou et al. (2023), proposent *clusterMLD*, une méthode de CAH basée sur une mesure de dissimilarité, coût de fusion, définie comme le coût de l'application d'unique B-splines sur les groupes fusionnés et sur les groupes pris séparément. Ce coût, analogue à celui de *Ward* entre deux groupes C_k et C_l, est donné par :

$$\mathcal{D}(C_k, C_l) = \frac{\mathrm{SSR}(C_{k,l}) - \mathrm{SSR}(C_k) - \mathrm{SSR}(C_l)}{d} \Bigg/ \frac{\mathrm{SSR}(C_k) + \mathrm{SSR}(C_l)}{n_k + n_l - 2d}, \tag{6}$$

où $\mathrm{SSR}(C_k)$ correspond à la somme des résidus au carré du modèle ajusté sur le groupe C_k, d est le nombre de fonctions de base, n_k et n_l sont les cardinaux des individus dans les deux groupes.

Dans la section suivante, nous présentons une étude comparative évaluant l'impact de l'intégration de la dérivée première pour les différentes méthodes de classification précédemment présentées.

4 Étude comparative sur données simulées

L'ensemble du code R permettant de reproduire l'étude est disponible via le lien suivant : `https://github.com/Diallo0/EGC-2026`

4.1 Plan de simulation

Données générées : On considère trois scénarios de génération de données fonctionnelles (Lu, 2024) :

Scénario I : $y_{ij} = f_{g(i)}(t_j) + \epsilon_{ij}$, où $f_{g(i)}(t_j)$ représente la courbe moyenne pour l'individu i dans le groupe g, $g(i)$ désigne le numéro de la classe de l'individu i. Plus précisément, on distinguera ces 4 courbes moyennes :

$$f_1(t) = 0.9\cos(2\pi t), \quad f_2(t) = 0.85\sin(4\pi t + 0.4),$$
$$f_3(t) = 0.95e^{-\frac{(t-0.3)^2}{2\times 0.07^2}} - 0.15, \quad f_4(t) = -0.9 + 1.8t.$$

Les individus du même groupe ont des courbes identiques. L'erreur aléatoire $\epsilon_{ij} \sim \mathcal{N}(0, 0.5^2)$ commune à tous les groupes.

Scénario II : $y_{ij} = f_{g(i)}(t_j) + r_i(t_j) + \epsilon_{ij}$, où l'effet aléatoire individuel est défini par : $r_i(t) = b_{i0} + b_{i1}t + b_{i2}t^2, \quad b_i = (b_{i0}, b_{i1}, b_{i2})^\top \sim \mathcal{N}(0, \Sigma)$,

$$\Sigma = \begin{pmatrix} 0.090 & 0.090 & -0.045 \\ 0.090 & 0.250 & -0.025 \\ -0.045 & -0.025 & 0.250 \end{pmatrix}.$$

Les effets aléatoires sont partagés par tous les groupes. Ainsi, les courbes des individus au sein du même groupe présentent de la variabilité, identique pour tous les groupes. La spécification de $f_{g(i)}(t_j)$ et ϵ_{ij} reste la même que dans le Scénario I.

Scénario III : $y_{ij} = f_{g(i)}(t_j) + r_{g(i)}(t_j) + \epsilon_{ij}$, où l'effet aléatoire est à la fois individuel et spécifique au groupe :

$$r_{g(i)}(t) = b_{g(i)0} + b_{g(i)1}t + b_{g(i)2}t^2, \quad b_{g(i)} = (b_{g(i)0}, b_{g(i)1}, b_{g(i)2})^\top \sim \mathcal{N}(0, \Sigma_g),$$

$$\Sigma_1 = \begin{pmatrix} 0.160 & 0.144 & -0.072 \\ 0.144 & 0.360 & -0.036 \\ -0.072 & -0.036 & 0.360 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 0.090 & 0.030 & -0.120 \\ 0.030 & 0.250 & 0.025 \\ -0.120 & 0.025 & 0.250 \end{pmatrix},$$
$$\Sigma_3 = \begin{pmatrix} 0.040 & -0.048 & 0.024 \\ -0.048 & 0.160 & 0.016 \\ 0.024 & 0.016 & 0.160 \end{pmatrix}, \Sigma_4 = \begin{pmatrix} 0.010 & -0.004 & 0.016 \\ -0.004 & 0.040 & -0.004 \\ 0.016 & -0.004 & 0.040 \end{pmatrix}.$$

Ici, la distribution des effets aléatoires dépend du groupe g. Ainsi, les courbes des individus dans un même groupe présentent de la variabilité, mais celle-ci peut différer entre groupes. La spécification de $f_{g(i)}(t_j)$ et ϵ_{ij} reste identique à celle du Scénario I.

Une illustration de ces différents scénarios de simulation est donnée sur la Figure 1, en Annexe, où chaque groupe est composé de $n_g = 100$ individus. Les courbes sont observées à intervalles réguliers sur l'intervalle $T = [0, 1]$, en $p = 20$ points de mesure équidistants (sans perte de généralité).

Méthodes comparées : Une fois les données simulées, elles sont lissées par MCP à l'aide de 50 fonctions de base de B-splines. Nous appliquons ensuite les différentes approches de classification présentées dans la section 3.2. La CAH avec la stratégie de *Ward* est appliquée à la matrice de dissimilarité obtenue grâce à la distance pondérée (4), notée $\tilde{D}_\omega$. On choisis K_0 et K_1 telle qu'on explique 99% de la variabilité des données. La même stratégie est appliquée à la matrice de dissimilarité calculée à partir de la distance pondérée fonctionnelle, notée D_ω. Enfin, elle est appliquée directement sur les coefficients issus du lissage, qui servira ici de référence (baseline).
Pour la méthode k-means, le nombre maximal d'itérations est fixé à 100 et le nombre d'initialisations aléatoires des centroïdes à 20. La partition retenue est celle qui minimise l'inertie intra.
Pour la méthode PAM, les médoïdes initiaux sont choisis aléatoirement parmi les courbes, puis l'algorithme est itéré jusqu'à ce qu'aucun changement de médoïde ne se produise.

Nous considérons également l'approche récente $SPFC_{\tilde{D}_\omega}$, avec ses paramètres d'initialisation standards, à savoir un choix de K_0 et K_1 afin d'expliquer respectivement 90% et 70% de la variabilité des données. Enfin, nous incluons la CAH basée sur les B-splines de Zhou et al. (2023), également avec ses paramètres standards. Cette dernière n'intègre pas la dérivée, reposant sur un coût de fusion, noté, $\mathcal{D}$ (6), semblable à celui de *Ward*, qui servira de référence.

Critère : Notre objectif est de mesurer l'impact de l'intégration de la dérivée première à l'aide d'un critère externe, l'Ajusted Rand Index (ARI). Il évalue la capacité à retrouver la partition de référence (connue dans le cadre de cette simulation) en comparant les partitions obtenues par les différentes approches à celle-ci. Ce critère est fourni sous forme de moyennes avec leurs écart-types, obtenues à partir d'une série de $B = 100$ répétitions.

4.2 Résultats

Scénario I : Dans le Tableau 1, les meilleurs résultats sont obtenus pour $\omega = 0$. L'accent est donc mis principalement sur l'amplitude des courbes. Ce résultat était attendu dans ce scénario, car dans notre simulation les courbes d'une classe ne diffèrent que par le bruit. Après lissage (débruitage), les courbes de moyenne différentes se distinguent mieux en amplitude, comme on peut le voir dans la Figure 1(a). L'ARI moyen dans ce cas varie entre 0.974 et 0.981 sauf pour la méthode $SPFC_{\tilde{D}_\omega}$ qui obtient un ARI moyen de 0.618. Les méthodes de références, baseline et la CAH fonctionnelle ($\mathcal{D}$), présentent des scores moyens d'ARI élevés, respectivement 0.977, 0.992, ce dernier étant le meilleur score dans ce scénario (voir Tableau 3).

Scénario II et III : Dans ces scénarios, l'introduction d'une variabilité intra-groupe (effets aléatoires r_i et $r_{g(i)}$) rend les courbes plus mélangées en amplitude à chaque instant, de sorte que la séparation repose sur des différences plus fines entre les courbes. Dans ce contexte, l'intégration d'un terme de la dérivé dans la distance ($0 < \omega \leq 1$) améliore nettement les performances des méthodes de classification. Comme on peut le

voir en gras dans le Tableau 1. Par exemple, dans le scénario II et pour les méthodes k-means, PAM et CAH (*Ward*), les ARI restent en dessous de 0.4 sans ce terme ($\omega = 0$), alors qu'ils atteignent un minimum de 0.651 et un maximum de 0.821 dès que la dérivée est prise en compte dans la distance. On constate le même effet dans le scénario III. Néanmoins, $SPFC_{\tilde{D}_\omega}$ donne de meilleurs résultats lorsque $\omega = 0$, et ce, dans tous les scénarios. Cela pourrait être dû au fait qu'il garde plus de variabilité sur les amplitudes de courbes que sur leur dynamique (sous ses initialisations standards). Ainsi, il se peut que des informations discriminantes soient portées par les composantes non retenues.
Dans le scénario II, les meilleurs ARI atteignent tout au plus 0.821 (plage entre 0.651 et 0.821) comme on peut le voir dans le Tableau 1 en gras. Par contre, pour le scénario III, les meilleurs ARI vont de 0.742 jusqu'à 0.935. Ainsi, les possibilités de retrouver la partition de référence sont significativement renforcées dans le scénario III grâce à la variabilité spécifique à chaque groupe. D'ailleurs on peut noter que la CAH fonctionnelle ($\mathcal{D}$) atteint un ARI de 0.884 (voir Tableau 3) pour le scénario III, alors qu'il a un faible ARI de 0.091 pour le scénario II.

Scénario	**Méthode**	**Distance**	$\omega = 0$	$\omega = 0.2$	$\omega = 0.4$	$\omega = 0.6$	$\omega = 0.8$	$\omega = 1$
Scénario I	k-means	$\boldsymbol{D}_\omega$	**0.981** (0.011)	0.981 (0.012)	0.981 (0.012)	0.980 (0.012)	0.978 (0.012)	0.670 (0.040)
		$\tilde{\boldsymbol{D}}_\omega$	**0.981** (0.011)	0.981 (0.012)	0.980 (0.011)	0.978 (0.012)	0.975 (0.013)	0.669 (0.026)
	PAM	$\boldsymbol{D}_\omega$	**0.974** (0.017)	0.974 (0.016)	0.974 (0.017)	0.974 (0.017)	0.973 (0.017)	0.710 (0.100)
		$\tilde{\boldsymbol{D}}_\omega$	**0.974** (0.017)	0.968 (0.017)	0.966 (0.016)	0.964 (0.016)	0.960 (0.017)	0.918 (0.034)
	CAH (*Ward*)	$\boldsymbol{D}_\omega$	0.977 (0.013)	**0.978** (0.013)	0.977 (0.013)	0.976 (0.014)	0.974 (0.015)	0.645 (0.055)
		$\tilde{\boldsymbol{D}}_\omega$	**0.977** (0.013)	0.976 (0.013)	0.974 (0.014)	0.963 (0.021)	0.903 (0.047)	0.654 (0.061)
	SPFC	$\tilde{\boldsymbol{D}}_\omega$	**0.618** (0.057)	0.468 (0.101)	0.446 (0.107)	0.438 (0.109)	0.435 (0.109)	0.433 (0.109)
Scénario II	k-means	$\boldsymbol{D}_\omega$	0.291 (0.090)	0.355 (0.083)	0.435 (0.063)	0.513 (0.045)	0.632 (0.069)	**0.692** (0.065)
		$\tilde{\boldsymbol{D}}_\omega$	0.297 (0.087)	0.490 (0.044)	0.599 (0.058)	0.716 (0.081)	**0.821** (0.062)	0.692 (0.065)
	PAM	$\boldsymbol{D}_\omega$	0.348 (0.076)	0.378 (0.072)	0.408 (0.074)	0.444 (0.071)	0.512 (0.083)	**0.707** (0.084)
		$\tilde{\boldsymbol{D}}_\omega$	0.352 (0.075)	0.579 (0.114)	0.714 (0.090)	0.784 (0.059)	0.808 (0.041)	**0.818** (0.043)
	CAH (*Ward*)	$\boldsymbol{D}_\omega$	0.335 (0.091)	0.367 (0.109)	0.429 (0.122)	0.519 (0.098)	0.581 (0.106)	**0.651** (0.026)
		$\tilde{\boldsymbol{D}}_\omega$	0.314 (0.089)	0.610 (0.130)	0.773 (0.103)	**0.807** (0.060)	0.779 (0.061)	0.676 (0.075)
	SPFC	$\tilde{\boldsymbol{D}}_\omega$	**0.712** (0.112)	0.487 (0.116)	0.447 (0.118)	0.432 (0.119)	0.425 (0.119)	0.420 (0.120)
Scénario III	k-means	$\boldsymbol{D}_\omega$	0.485 (0.039)	0.526 (0.072)	0.581 (0.075)	0.631 (0.109)	**0.847** (0.120)	0.678 (0.042)
		$\tilde{\boldsymbol{D}}_\omega$	0.485 (0.038)	0.593 (0.074)	0.757 (0.149)	**0.890** (0.069)	0.868 (0.079)	0.681 (0.051)
	PAM	$\boldsymbol{D}_\omega$	0.604 (0.122)	0.639 (0.117)	0.694 (0.113)	0.742 (0.110)	**0.793** (0.090)	0.689 (0.084)
		$\tilde{\boldsymbol{D}}_\omega$	0.611 (0.120)	0.835 (0.089)	0.888 (0.040)	**0.892** (0.036)	0.885 (0.029)	0.862 (0.043)
	CAH (*Ward*)	$\boldsymbol{D}_\omega$	0.594 (0.076)	0.653 (0.101)	0.679 (0.107)	0.683 (0.121)	**0.742** (0.161)	0.647 (0.028)
		$\tilde{\boldsymbol{D}}_\omega$	0.592 (0.077)	**0.935** (0.047)	0.929 (0.034)	0.878 (0.044)	0.833 (0.053)	0.651 (0.052)
	SPFC	$\tilde{\boldsymbol{D}}_\omega$	**0.667** (0.087)	0.493 (0.110)	0.458 (0.120)	0.445 (0.122)	0.439 (0.123)	0.433 (0.123)

TAB. 1 : ARI moyen (écart-type entre parenthèses) selon le scénario, la méthode et la distance utilisée, pour chaque valeur de ω. Les meilleures valeurs pour chaque méthode et scénario sont en gras.

On peut également étudier l'effet du terme de la dérivé selon les différentes valeurs de ω, sur la compacité intra-groupe et la séparabilité inter-groupes. Pour cela, nous utilisons l'indice Silhouette, et les résultats correspondants sont présentés dans le Tableau 2. On constate globalement qu'on a un meilleur indice Silhouette lorsque la distance qui donne la partition n'intègre pas un terme de la dérivée ($\omega = 0$, scénario I) ou exclusivement issue des dérivées des fonctions ($\omega = 1$, scénario II et III). Ce qui est conforme aux simulations, car en effet, dans les deux derniers scénarios, l'introduction des effets aléatoires rendent les courbes moins distinguables en amplitude, mais ceci est moins marqué sur les dérivées.

Scénario	Méthode	Distance	$\omega = 0$	$\omega = 0.2$	$\omega = 0.4$	$\omega = 0.6$	$\omega = 0.8$	$\omega = 1$
Scénario I	k-means	$\boldsymbol{D}_\omega$	**0.499** (0.009)	0.498 (0.009)	0.496 (0.010)	0.493 (0.010)	0.485 (0.012)	0.441 (0.028)
		$\tilde{\boldsymbol{D}}_\omega$	**0.503** (0.009)	0.295 (0.007)	0.278 (0.008)	0.258 (0.008)	0.232 (0.009)	0.347 (0.015)
	PAM	$\boldsymbol{D}_\omega$	**0.498** (0.010)	0.497 (0.010)	0.495 (0.010)	0.492 (0.011)	0.484 (0.012)	0.399 (0.055)
		$\tilde{\boldsymbol{D}}_\omega$	**0.502** (0.010)	0.294 (0.008)	0.277 (0.008)	0.256 (0.009)	0.230 (0.009)	0.269 (0.017)
	CAH (*Ward*)	$\boldsymbol{D}_\omega$	**0.498** (0.010)	0.497 (0.010)	0.495 (0.010)	0.492 (0.011)	0.484 (0.012)	0.443 (0.032)
		$\tilde{\boldsymbol{D}}_\omega$	**0.502** (0.010)	0.294 (0.008)	0.277 (0.008)	0.256 (0.009)	0.221 (0.013)	0.337 (0.027)
Scénario II	k-means	$\boldsymbol{D}_\omega$	0.246 (0.021)	0.252 (0.017)	0.262 (0.012)	0.267 (0.009)	0.271 (0.012)	**0.420** (0.046)
		$\tilde{\boldsymbol{D}}_\omega$	0.251 (0.020)	0.149 (0.006)	0.147 (0.008)	0.150 (0.011)	0.163 (0.011)	**0.323** (0.025)
	PAM	$\boldsymbol{D}_\omega$	0.246 (0.020)	0.248 (0.018)	0.248 (0.018)	0.252 (0.018)	0.260 (0.018)	**0.387** (0.050)
		$\tilde{\boldsymbol{D}}_\omega$	0.250 (0.020)	0.149 (0.007)	0.150 (0.008)	0.153 (0.009)	0.158 (0.010)	**0.252** (0.018)
	CAH (*Ward*)	$\boldsymbol{D}_\omega$	0.221 (0.019)	0.224 (0.020)	0.228 (0.017)	0.237 (0.017)	0.242 (0.019)	**0.431** (0.023)
		$\tilde{\boldsymbol{D}}_\omega$	0.222 (0.020)	0.141 (0.011)	0.146 (0.011)	0.146 (0.012)	0.143 (0.013)	**0.310** (0.036)
Scénario III	k-means	$\boldsymbol{D}_\omega$	0.302 (0.010)	0.301 (0.012)	0.306 (0.013)	0.308 (0.014)	0.299 (0.019)	**0.428** (0.039)
		$\tilde{\boldsymbol{D}}_\omega$	0.305 (0.010)	0.179 (0.008)	0.178 (0.008)	0.176 (0.008)	0.180 (0.011)	**0.330** (0.025)
	PAM	$\boldsymbol{D}_\omega$	0.286 (0.015)	0.283 (0.015)	0.285 (0.016)	0.289 (0.014)	0.289 (0.017)	**0.398** (0.048)
		$\tilde{\boldsymbol{D}}_\omega$	**0.288** (0.015)	0.174 (0.007)	0.173 (0.008)	0.173 (0.008)	0.173 (0.009)	0.260 (0.017)
	CAH (*Ward*)	$\boldsymbol{D}_\omega$	0.283 (0.015)	0.288 (0.016)	0.295 (0.017)	0.301 (0.017)	0.303 (0.019)	**0.431** (0.030)
		$\tilde{\boldsymbol{D}}_\omega$	0.287 (0.014)	0.173 (0.008)	0.170 (0.009)	0.163 (0.012)	0.158 (0.014)	**0.324** (0.023)

TAB. 2 : Indice Silhouette moyen (écart-type entre parenthèses) selon le scénario, la méthode et la distance utilisée, pour chaque valeur de ω. En gras, la plus grande valeur moyenne pour chaque méthode et scénario.

Méthode	Indice	Scénario I	Scénario II	Scénario III
$\mathcal{D}$	ARI	**0.992** (0.009)	0.091 (0.120)	**0.884** (0.037)
Baseline	ARI	0.977 (0.014)	0.338 (0.094)	0.587 (0.068)
Baseline	SIL	0.493 (0.010)	0.221 (0.022)	0.284 (0.012)

TAB. 3 : Moyennes des indices (écarts-types entre parenthèses) pour les méthodes de références sur chaque scénario.

5 Données réelles

Maintenant que l'intérêt de l'intégration de la dérivée première pour améliorer la qualité des partitions a été mis en évidence sur des données simulées, cette approche est évaluée sur des jeux de données réelles. Nous commencerons par les présenter, puis nous donnerons les résultats d'ARI des différentes méthodes de classification en fonction des pondérations ω.

Présentation : Les jeux de données, illustrés sur la Figure 2, sont couramment utilisés dans la littérature fonctionnelle. Nous considérons tout d'abord le jeu de données ***Phoneme***, qui contient les log-périodogrammes de cinq phonèmes anglais enregistrés. Il comprend 450 courbes spectrales, chacune mesurée en 256 points de fréquence, accompagnées de l'étiquette du phonème correspondant. Le jeu de données ***Tecator*** regroupe les spectres d'absorbance infrarouge de morceaux de viande mesurés sur 100 longueurs d'onde comprises entre 850 et 1050 nanomètres (nm). Il comporte 215 échantillons, chacun associé à trois variables de référence : la teneur en eau (*moisture*), en graisse (*fat*) et en protéines (*protein*). À partir de la variable *fat*, une labellisation binaire est obtenue selon un seuil fixé à 20%. Les courbes sont observées sur une grille régulière. Enfin, le jeu de données ***Growth*** contient les courbes de croissance (taille en cm) de 93 enfants (39 garçons et 54 filles), mesurées à 31 âges compris entre 1 et 18 ans.

Résultats : Pour les données du *Phonème*, l'utilisation des courbes reconstruites (coefficients estimés) seules donne un ARI élevé de 0.690, comme le montre le Tableau 5. Tout comme dans le scénario I sur les données simulées, après débruitage par le lissage, les courbes se distinguent bien en amplitude (voir Figure 2). Ce qui explique les valeurs d'ARI importantes pour ($\omega = 0$) pour les méthodes k-means, PAM et CAH (*Ward*) dans le Tableau 4 : une plage de valeurs entre 0.741 et 0.760. Cependant, pour *Tecator* et *Growth*, qui ressemblent davantage à nos scénarios II et III. Les courbes sont proches en amplitude et se distinguent davantage en termes de dynamique (voir Figure 2). On observe qu'intégrer l'information de la dérivée première améliore la capacité des méthodes à retrouver la véritable partition. En effet, pour les données *Tecator*, on peut noter que pour les méthodes k-means, PAM et CAH (*Ward*), les meilleures partitions sont obtenues sur la matrice de dissimilarité issue exclusivement de l'information de la dérivée ($\omega = 1$). On passe d'un ARI maximal de 0.142 pour $\omega = 0$ à un ARI minimal de 0.579 pour $\omega = 1$. Sur *Growth*, les meilleures partitions sont obtenues également en intégrant un terme de la dérivée. Par exemple, on obtient un ARI de 0.87 pour la méthode CAH (*Ward*) sur la matrice de dissimilarité donnée par $\tilde{D}_\omega$ pour $\omega = 0.6$ alors qu'on a un ARI de 0.01 pour $\omega = 0$. $SPFC_{\tilde{D}_\omega}$ échoue pour *Growth* de par son opération de réaffectation des courbes, qui vide un des deux groupes. La baseline et la CAH fonctionnelle ($\mathcal{D}$) ont du mal à retrouver la vraie partition pour *Tecator* et *Growth*. Elles font au mieux un ARI de 0.331.

Données	Méthode	Distance	$\omega = 0$	$\omega = 0.2$	$\omega = 0.4$	$\omega = 0.6$	$\omega = 0.8$	$\omega = 1$
Phoneme	k-means	D_ω	**0.741**	0.741	0.741	0.741	0.741	0.712
		$\tilde{D}_\omega$	0.741	0.764	0.759	**0.784**	0.723	0.729
	PAM	D_ω	0.746	0.746	0.746	**0.751**	0.735	0.528
		$\tilde{D}_\omega$	0.709	0.732	0.734	0.729	**0.735**	0.588
	CAH (*Ward*)	D_ω	0.760	0.700	**0.770**	0.710	0.520	0.370
		$\tilde{D}_\omega$	0.760	0.720	0.710	0.720	**0.800**	0.640
	SPFC	$\tilde{D}_\omega$	0.463	0.463	0.459	0.469	**0.471**	0.380
Tecator	k-means	D_ω	0.135	0.135	0.135	0.135	0.152	**0.640**
		$\tilde{D}_\omega$	0.135	0.175	0.249	0.342	0.613	**0.640**
	PAM	D_ω	0.142	0.150	0.157	0.197	0.241	**0.579**
		$\tilde{D}_\omega$	0.142	0.181	0.206	0.335	0.518	**0.738**
	CAH (*Ward*)	D_ω	0.140	0.080	0.080	0.270	0.190	**0.580**
		$\tilde{D}_\omega$	0.140	0.110	0.100	0.160	0.300	**0.620**
	SPFC	$\tilde{D}_\omega$	**0.276**	0.276	0.276	0.276	0.266	0.053
Growth	k-means	D_ω	0.074	0.087	0.165	**0.719**	0.612	0.612
		$\tilde{D}_\omega$	0.074	0.165	**0.719**	0.612	0.612	0.612
	PAM	D_ω	0.062	0.073	0.086	**0.612**	0.612	0.546
		$\tilde{D}_\omega$	0.115	0.100	**0.647**	0.514	0.546	0.546
	CAH (*Ward*)	D_ω	0.010	0.240	0.240	0.720	**0.830**	0.580
		$\tilde{D}_\omega$	0.010	0.160	0.760	**0.870**	0.790	0.580
	SPFC	$\tilde{D}_\omega$	-0.006	0.000	0.000	0.000	0.000	**0.000**

TAB. 4 : Valeurs d'ARI en fonction de ω pour les différentes méthodes intégrant l'information de la dérivée sur les trois jeux de données.

Méthode	Phoneme	Tecator	Growth
$\mathcal{D}$	0.155	0.027	0.331
Baseline	0.690	0.142	0.014

TAB. 5 : Valeurs d'ARI pour les méthodes de références sur les trois jeux de données.

6 Conclusions et perspectives

Ce travail souligne l'intérêt d'intégrer la dérivée première dans la classification de données fonctionnelles. La dérivée met en évidence des caractéristiques locales des courbes (changements de pente, points d'inflexion, etc.) pouvant être discriminantes. L'efficacité de cette approche repose sur un lissage adéquat via une pénalisation, équilibrant sur-ajustement et sous-ajustement.

Nous avons montré, par une étude de simulation et à travers des jeux de données réelles de la littérature, que l'utilisation de l'information issue de la dérivée améliore la détection de la vraie partition (simulée) lorsque les différences entre groupes tiennent aussi à des variations locales. Toutefois, les performances des méthodes dépendent de la nature des données : aucune n'est universellement optimale, il est donc nécessaire d'adapter la distance à la problématique étudiée.

Enfin, ce travail pourrait être poursuivi en intégrant la dérivée seconde, comme le suggèrent D'Urso et Vichi (1998); Ferraty et Vieu (2006), ce qui constituerait un prolongement naturel. En effet, celle-ci pourrait permettre d'améliorer la classification en introduisant, par l'accélération, une information discriminante.

Références

D'Urso, P. et M. Vichi (1998). *Dissimilarities between trajectories of a three-way longitudinal data set*. Berlin : Springer.

Ferraty, F. et P. Vieu (2006). *Nonparametric Functional Data Analysis*. France : Springer.

Kaufman, L. et P. J. Rousseeuw (2009). *Finding groups in data : an introduction to cluster analysis*. Hoboken, NJ : John Wiley & Sons.

Lu, Z. (2024). Clustering longitudinal data: A review of methods and software packages. *International Statistical Review*.

Meng, Y., J. Liang, F. Cao, et Y. He (2018). A new distance with derivative information for functional k-means clustering algorithm. *Information Sciences 463*, 166–185.

Ramsay, J. O. et B. W. Silverman (2005). *Functional Data Analysis*. Berlin: Springer.

Yu, P., G. Shi, C. Wang, et X. Song (2025). Distance-based clustering of functional data with derivative principal component analysis. *Journal of Computational and Graphical Statistics 34*(1), 47–58.

Zhou, J., Y. Zhang, et W. Tu (2023). clustermld: An efficient hierarchical clustering method for multivariate longitudinal data. *Journal of Computational and Graphical Statistics 32*(3), 1131–1144.

Annexe

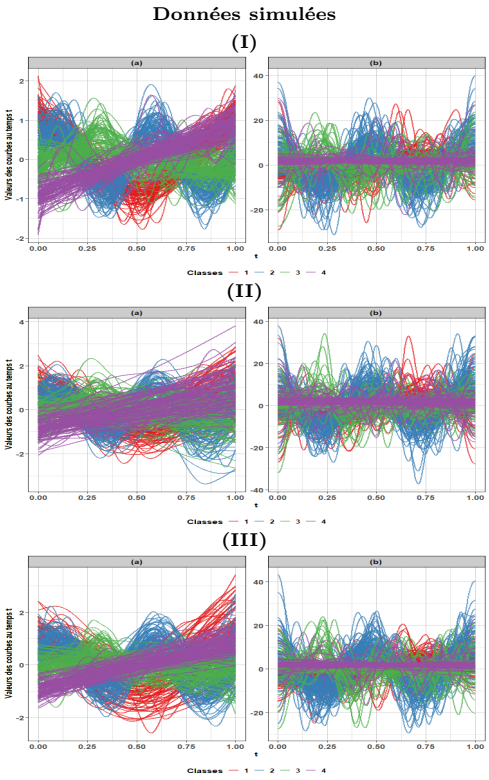

FIGURE 1 : **Scénarios : I, II, III :** courbes reconstruites par lissage MCP et leurs dérivées, évaluées sur une grille fine (1000 instants équidistants de $T = [0, 1]$).

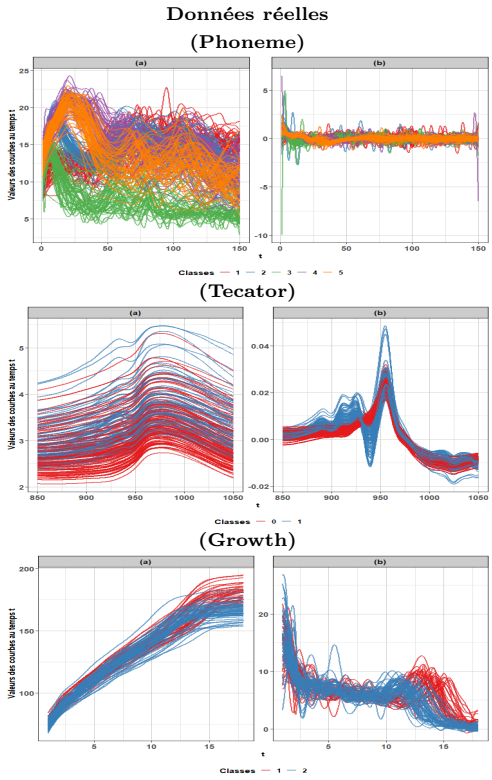

FIGURE 2 : **Données réelles :** Phoneme (1), Tecator (2), Growth (3) : courbes reconstruites par lissage MCP et leurs dérivées, évaluées sur une grille fine (1000 instants équidistants).

Summary

In functional data classification using distance-based approaches, the partition obtained is sensitive to the choice of distances. Several distance measures have been proposed in the literature. Here, we study the impact on the partition of integrating the first derivative into the calculation of distances in terms of internal (Silhouette) and external (adjusted Rand index) quality indices. To do this, we first conduct a comparative study of several of these approaches through different functional data simulation scenarios where the groups differ in the shape of the curves. This study is then supplemented by an evaluation based on different real data sets. The results show that the integration of the derivative can have a significant effect on the quality of the partition depending on the nature of the differences between groups.